2013 IEEE Security and Privacy Workshops

# Inside the SCAM Jungle: A Closer Look at 419 Scam Email Operations

Jelena Isacenkova*, Olivier Thonnard†, Andrei Costin*, Davide Balzarotti*, Aurelien Francillon*
*Eurecom, France
†Symantec Research Labs

***Abstract*—Nigerian scam is a popular form of fraud in which the fraudster tricks the victim into paying a certain amount of money under the promise of a future, larger payoff.**

**Using a public dataset, in this paper we study how these forms of scam campaigns are organized and evolve over time. In particular, we discuss the role of phone numbers as important identifiers to group messages together and depict the way scammers operate their campaigns. In fact, since the victim has to be able to contact the criminal, both email addresses and phone numbers need to be authentic and they are often unchanged and re-used for a long period of time. We also present in details several examples of Nigerian scam campaigns, some of which last for several years - representing them in a graphical way and discussing their characteristics.**

## I. Introduction

Nigerian scam, also called "419 scam" as a reference to the 419 section in the Nigerian penal code, has been a known problem for several decades. Originally, the scam phenomenon started by postal mail, and then evolved into a business run via fax first, and email later. The prosecution of such criminal activity is complicated [4] and can often be evaded by criminals. As a result, reports of such crime still appear in the social media and online communities, e.g. *419scam.org* [1], exist to mitigate the risk and help users to identify scam messages.

Nowadays, 419 scam is often perceived as a particular type of *spam*. However, while most of the spam is now sent mainly by botnets and by compromised machines in bulk quantities, Nigerian scam activities are still largely performed in a manual way. Moreover, the underlying business and operation models differ. Spammers trap their victims through engineering effort, whereas scammers rely on human factors: pity, greed and social engineering techniques. Scammers use very primitive tools (if any) compared with other form of spam where operations are often completely automated. Even though today 419 scam messages are eclipsed by the large amount of spam sent by botnets, they are still a problem that causes substantial financial losses for a number of victims all around the world.

A distinctive characteristic of email fraud is the communication channel set up to reach the victim: from this point of view, scammers tend to use emails and/or phone numbers as their main contacts [5], while other forms of spam are more likely to forward their victims to specific URLs. For instance, a previous study of spam campaigns [9] (in which scam was considered a subset of spam) indicates that 59% of spam messages contain a URL.

The traditional spam and scam (non-Nigerian) scenarios have been already thoroughly studied (e.g. [9], [3]). Costin et al. [5] describe the use of phone numbers in a number of malicious activities. The authors show that the phone numbers used by scammers are often active for a long period of time and are reused over and over in different emails, making them an attractive feature to link together scam messages and identify possible campaigns. In this work, we test this hypotesis by using phone numbers and othe email features to automatically detect and study scam campaigns in a public dataset. In particular, we apply a multi-dimensional clustering technique to group together similar messages to identify criminals and study their operations. To the best of our knowledge, this is the first in-depth study of 419 campaigns.

Our analysis identifies over 1,000 different campaigns and, for most of them, phone numbers represent the cornerstone that allows us to link the different pieces together. Our experiments also show that it is possible to identify macro-clusters, i.e. large groups of scam campaigns probably run by the same criminal groups.

The rest of the paper is organized as follows. We start by describing the scam dataset (Section III), to which we apply our cluster analysis technique to extract scam campaigns, and compare the usage of email addresses and phone numbers (Section IV). In Section V we focus on a number of individual campaigns to present their characteristics. Finally, we draw our conclusions in Section VI.

## II. Related work

Scammers employ various techniques to harvest money from ingenuous victims. Tive [14] introduces


DOI 10.1109/SPW.2013.15

the tricks of Nigerian fee fraud and the philosophy of tricksters behind. Stajano and Wilson [10] studied a number of scam techniques and showed the importance of security engineering operations. A brief summary of Nigerian scam schemes was presented by Buchanan and Grant [4] indicating that Internet growth has facilitated the spread of cyber fraud. They also emphasize the difficulties of adversary prosecution - one of the main reasons why Nigerian scam is still an issue today. A more recent work by Oboh et al. [8] discusses the same problem of prosecution in a more global context taking the Netherlands as an example.

Another work by Goa et al. [7] proposes an ontology model for scam 419 email text mining demonstrating high precision in detection. A work by Pathak et al. [9] analyses email spam campaigns sent by botnets, describing their patterns and characteristics. The authors also show that 15% of the spam messages contained a phone number. A recent patent has been published by Coomer [2] on a technique that detects scam and spam emails through phone number analysis. This is the first mentioning of phone numbers being used for identifying scam. Costin et al. [5] studied the role of phone numbers in various online fraud schemes and empirically demonstrated it's significance in 419 scam domain. Our work extends Costin's study by focusing on scam campaign characterization, and relies on phone numbers and email addresses used by scammers.

## III. Dataset

In this section we describe the dataset we used for analyzing 419 scam campaigns and provide some insights into the scam messages. There are various sources of scam often reported by users and aggregated afterwards by dedicated communities, forums, and other online activity groups. The data chosen for our analysis come from `419scam.org` - a 419 scam aggregator - as it provides a large set of preprocessed data: email bodies, headers, and some already extracted emails attributes, like the scam category and the phone numbers. We downloaded the data from s website for a period spanning from January 2009 until August 2012.

The resulting dataset consists of 36,761 *419 scam* messages with 11,768 unique phone numbers. The general statistics of the data are shown in Table I. A first thing to notice is that the number of messages is three times bigger than the number of phone numbers. We did not notice any significant bursts of scam messages (verified on a monthly basis) during the three year span, suggesting that the email messages were constantly distributed over time. It is also important to note that the dataset is mostly limited to the European and African regions (with also a few Asian samples), which is due to the way the website owners are collecting and classifying the data.

Table I: General statistics table

| Description | Numbers |
|---|---|
| Scam messages | 36,761 |
| Unique messages | 26,250 |
| Total email addresses | 112,961 |
| Unique email addresses | 34,723 |
| Total phone numbers | 41,320 |
| Total unique phone numbers | 11,768 |
| Number of countries | 12 |

Phone numbers can also be used to identify a geographical location, typically the country were the phone is registered. Although it does not prove the origin of the scam, it still references a country and provides a certain level of confidence in the message content to their victims. For example, receiving a new partnership offer from UK could seem strange if the phone contact has a Nigerian prefix. Moreover, as shown in a previous study [5], mobile phone numbers are precise in indicating the country of residence of the phone owner as few roaming cases were found. Therefore, the phone attribute is precise in indicating geographical origins and could reliably be used in the study of 419 scam.

We then look at the time during which emails and phones were advertised by scammers in scam messages. 71% of the email addresses in our dataset were used only during one day. The remaining were used for an average duration of 79 days each. Phone numbers have a longer longevity than email addresses: 51% of the phone numbers were used only for one day. The rest of phone numbers were used on average for 174 days (around 6 months). This is an important feature in our data clustering analysis.

Table II summarizes the phone number geographical distribution. UK numbers are twice as common as Nigerian, and three times more common than the ones from Benin, the third biggest group. Netherlands and Spain are the leading countries in Europe. Note that UK should be considered as a special case. As reported by *419scam.org* and Costin et al. [5], all UK phone numbers in this dataset belong to personal numbering services – services used for forwarding phone calls to other phone numbers and serving as a masking service of the real destination for the callee. In our dataset there are 44% of such phone numbers (all with UK prefix), another 44% are mobile phone numbers and 12% of fixed lines [5].

The dataset is also labeled with a scam category. Around 64% of the emails are assigned to the category "419 scam" (general scam category). Most of the remaining emails (24%) belong to "Fake lottery".

Table II: Phones by countries

| Country | Total phones | Total in % |
|---|---|---|
| United Kingdom | 4,499 | 43% |
| Nigeria | 3,121 | 30% |
| Benin | 1,448 | 14% |
| South Africa | 562 | 5% |
| Spain | 372 | 4% |
| Netherlands | 263 | 3% |
| Ivory Coast | 89 | 1% |
| China | 68 | 1% |
| Senegal | 47 | 0.5% |
| Togo | 11 | 0.1% |
| Indonesia | 1 | 0.01% |

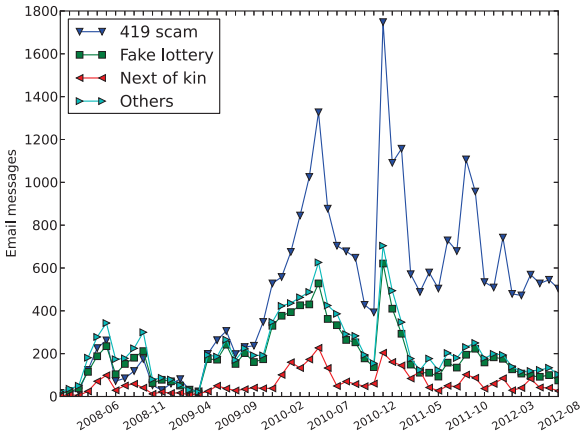



Figure 1: Scam email categories over time

However, this distribution has changed over time as shown in Figure 1. Especially, a big difference can be observed between 2009 and 2011, where in 2011 the "419 scam" became a dominant category. As of August 2012, there was 5 times more emails of "419 scam" than of "fake lottery" letters. This might be due to an outdated categorization process, as scam topics - like spam - may evolve over time. For this reason, in the next section we describe our process to automatically identify the scam topic based on the frequency of words in the messages. We also observe that most of the "fake lottery" scams are associated with European phone numbers, therefore suggesting a more targeted audience. In the majority of "419 scam" cases, scammers use African phone numbers with UK share being equivalent to Nigerian.

## IV. Data analysis

### A. Scam email clustering

To identify groups of scam emails that are likely part of a campaign orchestrated by the same group of people, we have clustered all scam messages using TRIAGE– a software framework for security data mining

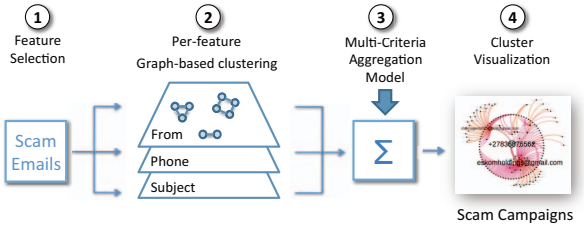



Figure 2: TRIAGE workflow example on scam dataset

that takes advantage of multi-criteria data analysis to group events based on subsets of common elements (*features*). Thanks to this multi-criteria clustering approach, TRIAGE identifies complex patterns in data, unveiling even *varying* relationships among series of connected or disparate events. TRIAGE is best described as a security tool designed for intelligence extraction helping to determine the patterns and behaviors of the intruders, and highlighting "how" they operate rather than "what" they do. The framework [11] has already demonstrated its utility in various analyses threats, *e.g.*, rogue AV campaigns [6], spam botnets [13] and targeted attacks [12].

Figure 2 illustrates the TRIAGE workflow, as applied to our scam data set. First, we select the email features, defined as decision criteria for linking the emails. In our experiment we used the sender email address (the *From*), email *subject*, *date*, *Reply-To* address, scammer *phone number* and email address found in the message body. Then, relationships among all email samples are built with respect to the selected features using appropriate comparison methods integrated in the framework. At the third step, the aggregation model fuses all features based on a set of weights defined to reflect feature importances and interactions during data fusion. We define parameters weighting thanks to the insights gained from previous study of scam phone numbers [5]. Hence, we assigned higher importance to phone, subject and reply address, and a lower importance to the email found in the body and the sending date.

As outcome, TRIAGE provides multi-dimensional clusters (MDC) of scam emails linked by at least a number of common traits. As explained in [11], the user can specify a threshold at which a link between clusters is created and that controls the relevance of the data within the same cluster. In our analysis, we choose a threshold of 0.30 by which any group of emails linked by a coalition of two similar features that includes at least the phone number, or by at least three similar features (no matter which combination), will exceed the threshold and thus create a cluster.

### B. Clustering results

We identified 1,040 clusters with TRIAGE that consist of at least 5 correlated scam messages. Because of the multi-criteria aggregation, we hypothesize that these clusters are quite likely reflecting different *scam campaigns* organized by the same individuals – as emails within the same clusters share several common traits. These, though, give no indication on the actual number of individuals that are behind each campaign. Based on the topologies of those campaigns, we anticipate there could be more than a single person in most cases. We look at this aspect in more details in section V.

Table III: Global statistics for the top 250 clusters

| Statistic | Average | Median | Maximum |
|---|---|---|---|
| Nr emails | 38 | 28 | 376 |
| Nr from | 13.9 | 9 | 181 |
| Nr reply | 6.2 | 5 | 56 |
| Nr subjects | 9.9 | 7 | 114 |
| Nr phones | 2.5 | 2 | 34 |
| Duration (in days) | 396 | 340 | 1,454 |
| Nr dates (distinct) | 27.9 | 22 | 259 |
| Compactness | 2.5 | 2.4 | 5.0 |

Table III provides some global statistics computed across the top-250 largest scam campaigns. In over half of these campaigns, scammers are using only two distinct phone numbers, but they still make use of more than 5 different mailboxes to get the answers from their victims. Most scam campaigns are rather *long-lived* (lasting on average about a year). We note that cluster sizes are small on average indicating that there are many small, isolated campaigns and only a few dozens of messages belong to the same campaign. This might be also an artefact of the data collection process; nevertheless, we anticipate that this could also reflect the scammers' behavior who may want to stay untraceable "by the radar". Indeed, bulk amounts of the same emails would have more potential to compromise their scamming operations, as this would become too visible to content-based scam filters and, hence, would get blocked on earlier stages of email filtering.

To confirm our intuition about the importance of certain features (phone numbers, and to a lesser extent, email addresses) and their effective role in identifying campaigns, we look at all similarity links within clusters. We observe that the features mainly responsible for linking scam messages in the clusters involve phone numbers (in 88% cases), followed by the *reply* email address (for 66% of the links). Not surprisingly, the *from* address (which can be easily spoofed) changes much more often and is used as linking feature in only 46% of clusters.

One could wonder about the longevity of these features, hence we also looked at phone numbers and email addresses from a time perspective. Figure 3 represents the usage of the same email addresses and phone numbers over time. The Y axis is density of the features that indicates their distribution in time on a 100% scale. As mentioned before, many of them are used for only one day, so there is a slight concentration on the left side of the plot. However, the phone numbers are more often reused over time than email addresses. This could be explained by an easy access to new mailboxes offered by many free email providers. As for the phone, they probably still require some financial investment compared with emails. We checked the domain names of email addresses used in our scam dataset and found that top 100 belong to webmail providers from all over the world. This finding suggests that email messages sent from such accounts would overpass sender-based anti-spam techniques are widely deployed today.

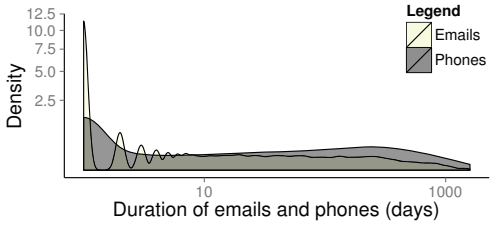



Figure 3: Duration of phone numbers and emails used by scammers, in days

### C. Content categorization

*419scam.org* [1], as mentioned, also categorizes the scam emails into 10 categories. We presented their shares in the dataset section III. Since this provided categorization is too broad, we decided to evaluate ourselves the categories in our dataset by measureing the word frequency in the body of the scam messages. To extract some more generalized knowledge of the clustered data, we create a list of the most repetitive keywords (after removing all the stop words) and group them into meaningful categories. As a result, we identified three big categories within clusters: money transfer and bank related fraud (54%), lottery scam (22%), and fake delivery services (11%). The rest is uncategorized and refers to 13% of the clusters. The repartition is similar to the one provided by the data source, except that the delivery services are separated into a separate category. The so called general "419 scam" category corresponds to letters about lost bank payments, compensations, and investment proposals. We grouped them together as they are very difficult to separate due to a number of keywords in common.

## V. Characterization of campaigns

This section provides deeper insights into 419 scam campaign orchestration. We present a few typical scam campaigns and we show connections between clusters, possibly run by the same group of scammers.

### A. Scam campaign examples

Figures 4, 5, 6 show examples of scam campaigns identified by TRIAGE, depicted with graph visualization tools developed in the VIS-SENSE project[1]. Those graphs are drawn with a circular layout that represents the various dates on which scam messages were sent. The dates are laid out starting from 9 o'clock (far left in the graph) and growing clockwise. Then, the cluster nodes are drawn with a force-directed placement algorithm. The big nodes on the graphs are mostly phone numbers and *From* email addresses. Smaller nodes represent mostly subjects and email addresses found in the *Reply-To* header or the message content.

Figure 4 is an example of a campaign impersonating a private company in South Africa, *ESKOM Holdings*. The ESKOM campaign was initially a fake lottery scam (left upper corner of Figure 4), but later switched to a different scam, while still re-using the same phone number. A noteworthy aspect of this campaign, shared with some other campaigns we found, is that it relies on few *From* emails addresses (i.e., the bigger nodes in the figure). The other email addresses are used with larger number of emails and change over time.

Another campaign, presented in Figure 5b, illustrates the roles of email addresses and phone numbers in *419 scam* over time. This campaign, that lasted for 1,5 year, changed topic over time (every 1 to 2 months), which is clearly visible by looking at the larger subgroups placed around the circle. These shorter campaigns were most probably run by the same scammers. We see that they almost completely changed the email addresses between different scam runs, but kept the same phone number. The email addresses were often selected to match the campaign topic and subjects.

Unfortunately, graphical interpretation of the campaigns is not always straightforward, as can be seen on Figure 5a. This graph was generated from a cluster of a recent campaign of iPhone-related scams that lasted for 1,5 years. The communication infrastructure of these scammers is much more diverse. The campaign relies on a large number of "disposable email addresses" that are seldom used for a long period. As opposed to previous examples, however, same or very similar subjects are often reused.

[1]The VIS-SENSE project: http://www.vis-sense.eu

### B. Macro clusters: connecting sub-campaigns

To try to find a connection between different campaigns, we searched for weaker connections between clusters. The goal was to pinpoint possibly larger-scale campaigns, which are made of loosely inter-connected scam operations (*i.e.* different scam *runs*). For this purpose we rely on email addresses and phone numbers as other attributes are less personal. We identify clusters that share email addresses and/or phone numbers, and use this information to build *macro-clusters*. We identify 845 isolated and 195 connected clusters, where the latter consists of 62 macro-clusters. The characteristics of top 6 macro-campaigns are shown in Table IV. These clusters are particularly interesting as they consists of a set of scam campaigns that appear to be interconnected and therefore could be orchestrated by the same people. Such macro-clusters span through time with bursts of different campaigns, topics and countries.

An example of such macro-campaigns is illustrated in Figure 6. This macro-cluster consists of 6 scam campaigns of various size that include UK and Nigerian phone numbers. We can distinguish them in the graph as they appear as groups with one or two bigger nodes (phone numbers) with a tail of connected nodes (email addresses). We notice that campaigns in this case are well separated by phones and emails dedicated for each campaign (or operation), and that there are only few overlaps over time. However, there is a small node just in the center that indicates their interconnection. Some contact details were reused and we used that for their correlation. These campaigns together lasted for 3,5 years. Over this time period, scammers have sent emails using 51 distinct subjects and 8 different phone numbers. In conclusion, we could describe this macro-campaign as run by a group of individuals from Nigeria that are changing contact details for each campaign and work with several scam categories. The topics diversity may suggest there might be a competition among scammers as they try to cover different online trick schemes instead of specializing in a single one.

### C. Geographical distribution of campaigns

Figure 8 shows the country distribution of the top 6 macro-campaigns. The last three campaigns are based in Africa and located in one or two countries. The first three are more Europe-oriented with some connections to Nigeria and Benin. These groups are competing in "fake lottery" scam, with the second group leading the pack and covering most of the countries. In comparison to previous similar study of scam campaign geographical distribution [5], we note that we encounter less UK and Nigerian numbers in campaigns, but at the same time confirm that scam campaigns can be multi-continental. The largest macro-campaign (#2) we

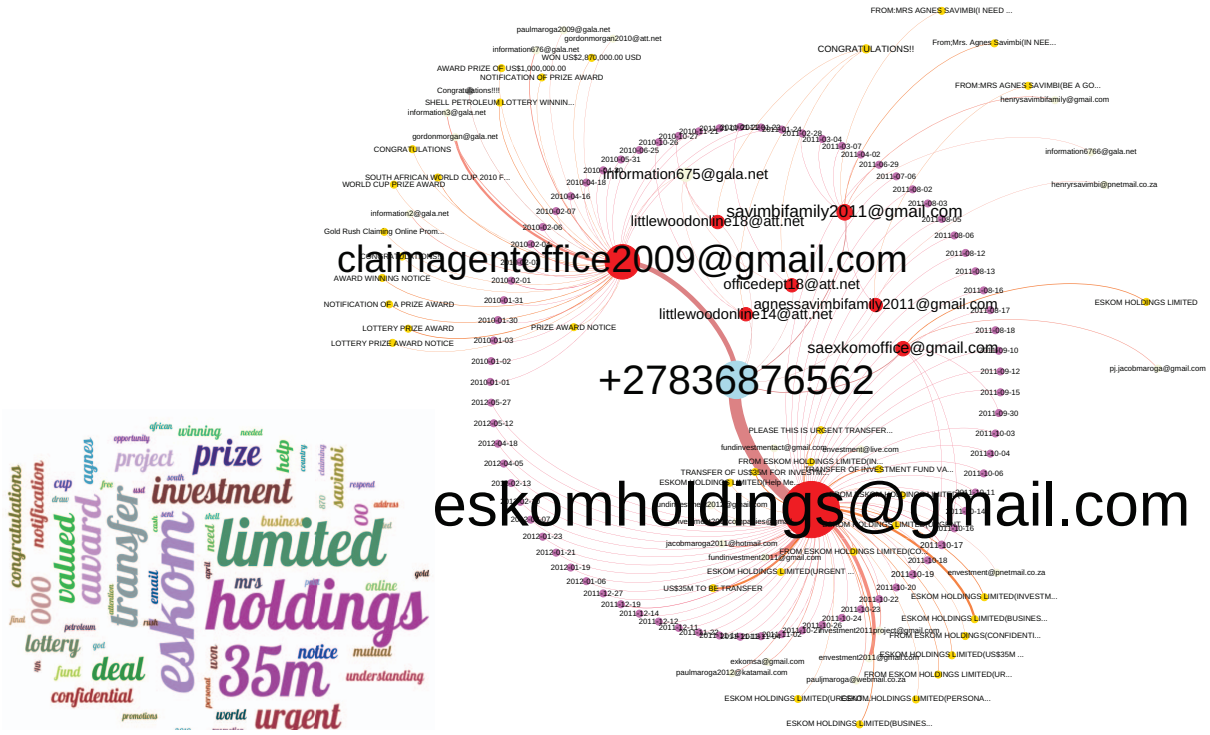
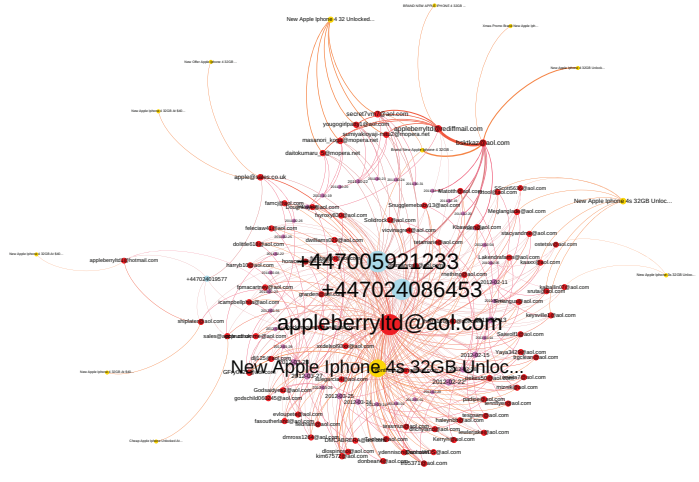


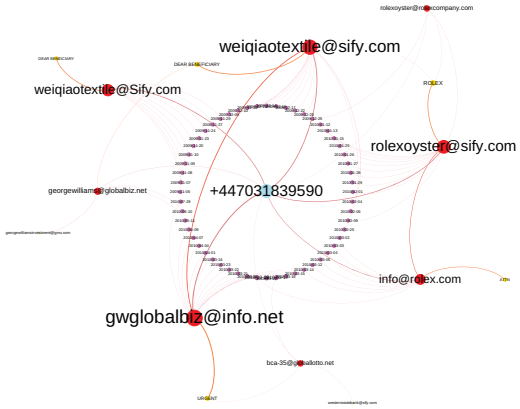




Figure 4: Lotteries (between 9 and 12 o'clock) and *ESKOM Holdings* impersonation.




(a) Very diverse iPhone scam campaign.




(b) Distinct sub-campaigns, connected through phone number.

Figure 5: Examples of other scam campaign structures.

Table IV: Macro-clusters, mean values of attributes

| Macro-cluster | Nr. of campaigns | Phones | Mailboxes | Subjects | Duration | Countries | Topics |
|---|---|---|---|---|---|---|---|
| 1 | 14 | 44 | 677 | 223 | 4 years | 4 | Lottery, lost funds, investments |
| 2 | 43 | 163 | 1,127 | 463 | 4 years | 7 | Lottery, banks, diplomats, FBI |
| 3 | 6 | 18 | 128 | 80 | 4 years | 4 | Lottery |
| 4 | 5 | 8 | 111 | 51 | 3,5 years | 2 | Packaging, Guiness lottery, loans |
| 5 | 6 | 7 | 201 | 96 | 1 year | 1 | Microsoft lottery, UPS & WU delivery, lost funds |
| 6 | 4 | 7 | 82 | 33 | 2 years | 1 | Lottery, lost payments |

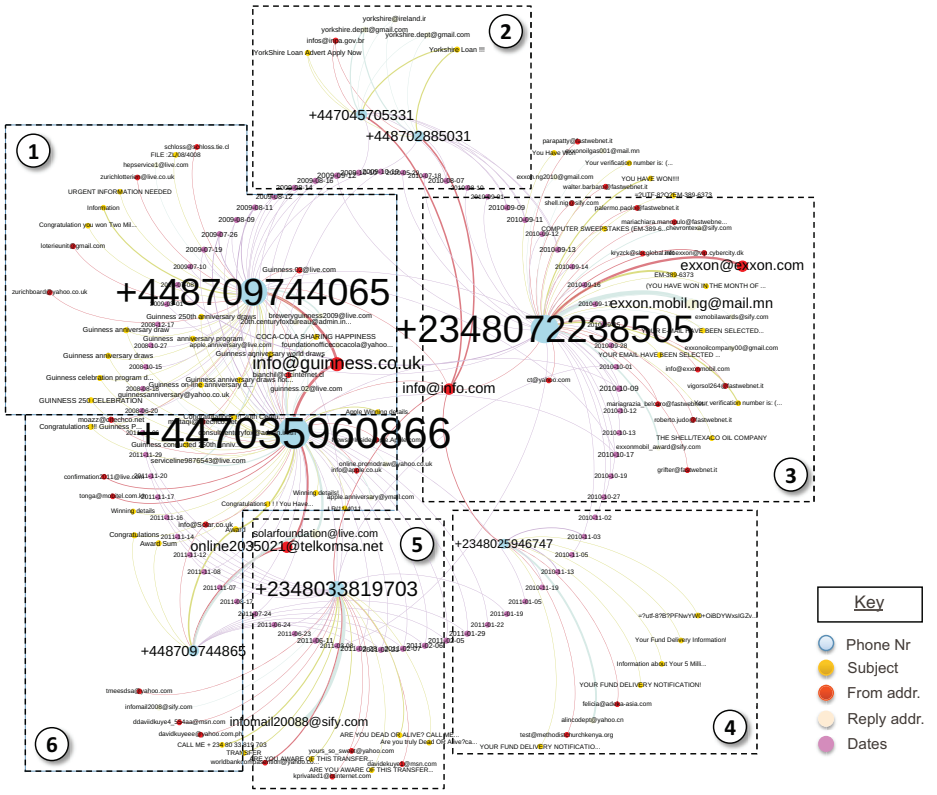



Figure 6: An example of macro-cluster. The nodes laid in clock-wise fashion reflect the timeline of the campaigns.

identified is potentially orchestrated by several groups of people distributed in several countries (based on the previous finding that mobile phones are rarely used outside its originating country).

To better understand how scammers are geographically located and how they work, we plot the number of emails per country for different datasets in Figure 7, for all data, clusters and macro-clusters respectively. We note also that the unclustered data is concentrated in African countries and that emails with a reference to European countries mostly get clustered with a quite big share in macro-clusters. This suggests that some scammers from African countries try to run more complicated Europe-based campaigns (mostly "fake lottery"). Those would possibly provide better revenues or provide more attractive (rich) victims. Additionally, as we saw

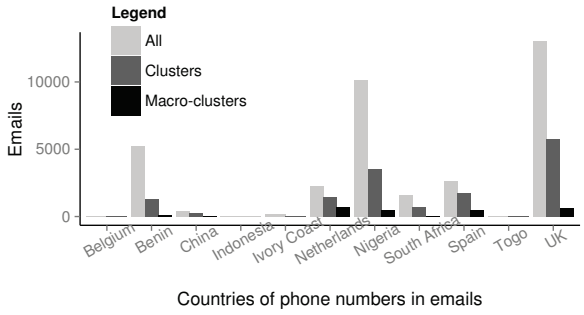



Figure 7: Largest macro-clusters distribution in countries.

in Figure 7, non-African emails seem to be often accompanied by African emails. Still, the majority of the other

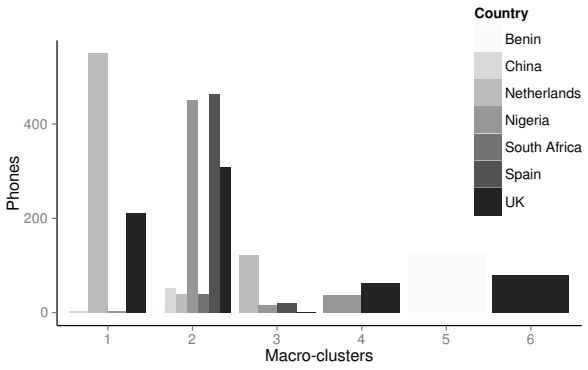



Figure 8: Country distribution in clustered data.

unclustered emails are from African continent, probably some campaigns run by a single person, possibly trying to start a new or improve an existing scam-business, and are thus more difficult to correlate and group into bigger scam clusters.

## VI. Conclusions

In this study, we identified around a thousand *419 scam* campaigns with the help of a multi-dimensional clustering technique for grouping similar emails. We showed that orchestration of such campaigns differs from traditional spam campaigns sent through botnets. Our analysis has unveiled a high diversity in scam orchestration methods, showing that scammer(s) can work on various topics within a campaign, thus probably competing with each other over trendy scam topics.

We also discussed the crucial role played by email addresses and phone numbers in scam business, in contrast with other cyber crime schemes where email addresses may be often spoofed and phone numbers rarely used. We also discovered that scammers re-use the same phone numbers and email addresses over long periods of time – sometimes up to 3 or 4 years. At the same time, scammers seem to send very low volumes of emails compared to spammers.

Finally, we uncovered the existence of *macro-campaigns*, groups of loosely linked together campaigns that are probably run by the same people. We found that some of these macro-campaigns are geographically spread over several countries, both African and European. We believe that our methods and findings could be leveraged to improve investigations of various crime schemes – other than scam campaigns as well.

## Acknowledgments

This research was partly supported by the European Commission's Seventh Framework Programme (FP7 2007-2013) under grant agreement nr. 257495 (VIS-SENSE). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Commission.

## References

[1] 419 Scam Fake Lottery Fraud Phone Directory. http://www.419scam.org/419-by-phone.htm.

[2] Patent US7917655: Method and system for employing phone number analysis to detect and prevent spam and e-mail scams. http://www.patentlens.net/patentlens/patent/US_7917655/en/.

[3] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker. *Spamscatter: Characterizing internet scam hosting infrastructure*. PhD thesis, University of California, San Diego, 2007.

[4] J. Buchanan and A. J. Grant. Investigating and Prosecuting Nigerian Fraud. *High Tech and Investment Fraud*, 2001.

[5] A. Costin, J. Isacenkova, M. Balduzzi, A. Francillon, and D. Balzarotti. The role of phone numbers in understanding cyber-crime schemes. *Technical Report, EURECOM*, RR-13-277, Feb 2013.

[6] M. Cova, C. Leita, O. Thonnard, A. D. Keromytis, and M. Dacier. An Analysis of Rogue AV Campaigns. In *Proceedings of the 13th international conference on Recent advances in intrusion detection*, RAID'10, pages 442–463, Berlin, Heidelberg, 2010. Springer-Verlag.

[7] Y. Gao and G. Zhao. Knowledge-based information extraction: a case study of recognizing emails of nigerian frauds. NLDB'05, pages 161–172, Berlin, Heidelberg, 2005. Springer-Verlag.

[8] J. Oboh and Y. Schoenmakers. Nigerian advance fee fraud in transnational perspective. *Policing multiple communities*, (15):235, 2010.

[9] A. Pathak, F. Qian, Y. C. Hu, Z. M. Mao, and S. Ranjan. Botnet spam campaigns can be long lasting: Evidence, implications, and analysis. *SIGMETRICS*, pages 13–24, 2009.

[10] F. Stajano and P. Wilson. Understanding scam victims: seven principles for systems security. *Commun. ACM*, 54(3):70–75, Mar. 2011.

[11] O. Thonnard. *A multi-criteria clustering approach to support attack attribution in cyberspace.* PhD thesis, École Doctorale d'Informatique, Télécommunications et Électronique de Paris, March 2010.

[12] O. Thonnard, L. Bilge, G. O'Gorman, S. Kiernan, and M. Lee. Industrial espionage and targeted attacks: Understanding the characteristics of an escalating threat. In *RAID*, pages 64–85, 2012.

[13] O. Thonnard and M. Dacier. A strategic analysis of spam botnets operations. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 162–171, New York, NY, USA, 2011. ACM.

[14] C. Tive. *419 scam: Exploits of the Nigerian con man*. iUniverse, 2006.